

DOCUMENT RESUME

ED 247 297

TM 840 478

AUTHOR Loyd, Brenda
TITLE The Effect of Number of Ability Intervals on the Stability of Item Bias Detection.
PUB DATE Feb 83
NOTE 19p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Baltimore, MD, February 24-27, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Ability; Black Students; College Entrance Examinations; English; High Schools; Latent Trait Theory; *Racial Bias; Reliability; *Statistical Analysis; *Test Bias; *Test Items; White Students
IDENTIFIERS American College Testing Program; *Chi Square; Kappa Coefficient

ABSTRACT

The chi-square procedure has been suggested as a viable index of test bias because it provides the best agreement with the three parameter item characteristic curve without the large sample requirement, computer complexity, and cost. This study examines the effect of using different numbers of ability intervals on the reliability of chi-square statistics as indicators of item bias in educational tests. In 1980 the English Usage exam from the American College Testing Program was administered to 922 black and 4423 white high school juniors and seniors at 16 test sites. Four randomly equivalent samples of black and white examinees were created. Three sets of ability grouping intervals, based on total raw score on the 75 items, were established using three, four, and five intervals. The same number of examinees were represented in each interval. A chi square index for each test item and three measures of stability (correlations of chi-square values; correlation of 0, 1 bias decisions; and the kappa index) were calculated. Results confirm previous studies and suggest that the chi-square index is fairly unreliable regardless of the number of grouping intervals and cannot lead to consistent decisions about item bias. (BS)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED247297

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

X This document has been reproduced as
received from the person or organization
originating it.
Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

THE EFFECT OF NUMBER OF ABILITY INTERVALS
ON THE STABILITY OF ITEM BIAS DETECTION

Brenda Loyd
University of Virginia

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

B. H. Loyd

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the Eastern Educational
Research Association in Baltimore, February, 1983

TM 840 478

THE EFFECT OF NUMBER OF ABILITY INTERVALS ON THE STABILITY OF ITEM BIAS DETECTION

INTRODUCTION

In recent years there has been a growing interest in exploring bias in educational tests. Educational tests are constructed to measure a particular underlying ability. When individuals from different groups (cultural, racial, sexual, etc.) with the same level of ability have different probabilities of success on a particular test, the test is said to be biased.

A more global definition acknowledges test bias as a complex interaction among what a test is intended to measure, how it measures, and the uses made of the resulting measurement. With such a definition bias could only be verified with the aid of external criteria used to assess the validity of a particular use of the test. Although this approach to bias is more encompassing, often such external criteria are not easily obtainable or do not exist. In addition, there is often a need to measure possible bias prior to the completion of the final form of a test in order to have the option of removing it from the test.

In the absence of external criteria, researchers have focused on the study of item bias. Many statistical methods have been developed to detect item bias. These methods differ in their theoretical bases, statistical complexity, and cost in terms of sample size, computer use, and test developer's time. The statistical methods range from the most complex techniques, employing the three-parameter item characteristic curve, to

the simpler techniques employing transformed item difficulties and chi-square statistics.

Several studies comparing item bias detection techniques have been conducted (Ironson & Subkoviak, 1979; Rudner, Getson, & Knight, 1980; and Shepard, Camilli, and Averill, 1980). These studies have attempted to determine which of several detection methods were the most useful and valid, and to examine the amount of agreement among the methods. Agreement among the methods was assessed in terms of correlations among the bias indices as well as whether the methods consistently identified the same items and the same number of items as biased. The three-parameter item characteristic curve (ICC) method is presented as the strongest method theoretically because its indices are less influenced by differences in group means and by confounding of item difficulty and item discrimination parameters (Lord, 1977). Empirical studies (Ironson, et al., 1979; and Shepard, et al., 1980) support the soundness of the three-parameter ICC method. However, the three-parameter method requires samples of at least 1000 and computer costs for the method are substantial. A bias detection technique using chi-square statistics was determined to agree with the three-parameter ICC model and was suggested as a reasonable substitute for the more costly method.

Studies which included information on the reliability of item bias indices across independent sources found only modest agreement (Scheuneman, 1980; and Linn, et al., 1981). Studies (Kolen and Hoover, 1982; and Loyd, 1982) which examined directly the internal consistency or stability of item bias methods reported surprisingly low reliabilities for bias indices across independent samples.

One explanation offered for the low reliability or lack of stability of the chi-square indices as reported by Loyd (1982) and Kolen and Hoover (1982) was that the recommended expected frequency per cell was not maintained for the sample size and the number of ability intervals selected for the studies. The bias detection technique employing a chi-square statistic has been suggested as a viable index because it provides the best agreement with the more theoretically sound three-parameter ICC model and yet can be implemented without the large sample requirement, the large computer complexity, or large cost; however, the lack of reliability or stability of the method deserves further examination. Therefore, the purpose of this study is to examine the effect of using different numbers of ability intervals on the stability of the chi-square item bias indices.

Chi-Square Method

A method of bias detection using a chi-square statistic was proposed by Scheuneman (1979) as a rough approximation of latent trait methods. Although there is no model of a latent ability, the total test score distribution is assumed to represent underlying ability and group means are roughly equated by creating matched score intervals on the total score distribution. The total score range is divided into a number of discrete intervals. In specifying the intervals, it is important to ensure that for each item the probability of answering the item correctly is less than one and greater than zero for each group. Expected proportions are calculated assuming ethnic group and probability of a correct response to be independent. Then the observed proportions are compared with the expected proportions using a goodness-of-fit chi-square statistic. Deviations from expected frequencies of the form $\frac{(O - E)^2}{E}$ are summed across intervals.

Large deviations result in large summed values which indicate a biased item.

Camilli (1979) recommended a modification of the chi-square procedure suggested by Scheuneman which includes both correct and incorrect responses in the fitting of expected values. The chi-square statistic for each item is calculated for each ethnic group by correct and incorrect responses (a 2 x 2 table) and then summed across the ability grouping intervals. If all items are uniformly difficult, the two chi-square methods agree exactly. With easier items, the Camilli method will find bias where the Scheuneman method will not. The Camilli method was the technique selected for use in this study.

The first step in using the chi-square procedure is to divide the total score scale into ability intervals. To determine the number of intervals and whether the intervals should cover approximately the same score width in terms of the total score scale or approximately the same number of people (ethnic groups combined), the following criteria are suggested:

1. expected frequencies should be at least five per cell, and
2. there should be a minimum number of observed responses in each cell (ten is suggested).

Once J is established (J = the number of ability intervals), the chi-square statistic can be calculated for each item by determining the chi-square value for each 2 x 2 table of ethnic group and correct/incorrect response. The final chi-square value is found by summing these values across the J ability grouping intervals.

One additional advantage of the chi-square method over other bias detection methods is that a statistical significance test may be used to establish the cutoff for signifying a biased item. (Other procedures employ even more arbitrary rules to decide how much bias is serious.) After the final chi-square value is calculated, it then can be compared with the theoretical chi-square distribution with J degrees of freedom.

METHODS

Test Materials

The test used in this study was the English Usage exam from the American College Testing Program battery of college entrance exams. The test is composed of 75 items which assess an examinee's understanding of the conventions of standard written English including punctuation, grammar, sentence structure, diction, style and logic. The test was administered to high school juniors and seniors in the fall of 1980.

Examinees

The examinees included in this study were selected from a national sample by first determining test sites that tested ten or more black students and ten or more white students. From these designated test sites, 16 sites were randomly selected. These 16 sites included 922 black examinees and 4425 white examinees.

For the first set of comparisons, black and white examinees were randomly assigned to one of four subsamples. Each subsample contained approximately 450 to 500 students; approximately half were black and half were white. Thus four randomly equivalent samples of black and white examinees were created.

Although there were nearly five times as many white examinees as black examinees in the total sample, it was essential that the number of blacks and whites be roughly the same in each subsample used in the study, because of the spurious bias indications produced by the chi-square method when the groups of interest are of widely varying size.

The relatively small sample size in this study was designed to reflect actual conditions under which bias detection procedures are normally used. These procedures are usually applied during "trial runs" when a new test is being checked by administering it to a relatively small sample. In such a situation, it would be unlikely that more than 250 black examinees would be included in the sample available.

Procedure

For the chi-square bias method, ability grouping intervals were created based on total raw score on the 75 items. Three sets of ability intervals were established, using three, four, and five intervals respectively. The intervals were established based on the criteria of minimum expected proportion per cell and so that approximately the same number of examinees (both black and white combined) were represented in each interval. For the first set of ability intervals, three ability groups were established, composed of those scoring

1. below 33,
2. from 33 to 42,
3. above 42.

For a second set of ability intervals, four ability groups were established, composed of those scoring

1. below 31,
2. from 31 to 37,
3. from 38 to 45,
4. above 45.

For a third set of ability intervals, five ability groups were established, composed of those scoring

1. below 26,
2. from 26 to 35,
3. from 36 to 40,
4. from 41 to 50,
5. above 50.

A chi-square index was calculated as described above for each item based on a frequency table of correct and incorrect responses by group (black and white) and summed across the ability intervals. Four estimates of the item bias index were computed for each of the three sets of ability intervals (one estimate from each subsample).

Pearson product moment correlations among the four estimates for each set of ability intervals were calculated.

Based on an α level of .05, statistical tests were conducted for each item under each condition. A critical value of 7.8 was established for item bias indices based on three ability groupings (.05 level for a chi-square distribution with three degrees of freedom). A critical value of 9.5 was established for item bias indices based on four ability groupings (.05 level for a chi-square distribution with four degrees of freedom). A critical value of 11.1 was established for item bias indices based on five ability groupings (.05 level for a chi-square distribution with five degrees

of freedom). If the item index exceeded the critical value the item was labeled biased. If the item failed to reach the critical value then the item was assumed to be unbiased.

Biased items were assigned a value of 1 and unbiased items were assigned a value of 0. Pearson product moment correlations among the four independent decisions were calculated for each of the three sets of ability intervals.

Because the Pearson product moment correlation underestimates the relationship for dichotomous variables, an additional index of stability or reliability was computed based on Kappa, κ , (Cohen, 1960), an index of agreement for use with nominal scales in which

$$\kappa = \frac{p_o - p_c}{1 - p_c}$$

where p_o is the observed proportion of agreement and p_c is the proportion of agreement expected by chance.

RESULTS

The mean, standard deviation, and sample size for each race group by each subsample are presented in Table 1. The means were consistently higher for whites than for blacks and the standard deviations were generally larger for whites than for blacks.

Insert Table 1 about here

The correlations among the four subsamples of the item chi-square indices for each grouping size are presented in Table 2. The correlations for the three interval grouping range from -.13 to .38. For the four interval grouping, the correlations range from -.16 to .31. For the five interval

grouping, the correlations range from $-.07$ to $.20$.

Insert Table 2 about here.

The correlations among the four subsamples of the bias decisions (coded 0, 1) are presented in Table 3 by number of intervals grouping. For the three interval grouping, correlations ranged from $.02$ to $.23$. Correlations for the four interval grouping ranged from $-.08$ to $.19$. Correlations for the five interval grouping ranged from $-.13$ to $.22$.

Insert Table 3 about here.

The kappa values calculated for each pair of subsample on the bias decisions are presented in Table 4 by number of ability groupings. Kappa ranges from $-.17$ to $.21$ for the three ability grouping. For the four ability grouping, kappa ranges from $-.08$ to $.18$. For the five ability grouping, kappa ranges from $-.12$ to $.19$.

Insert Table 4 about here.

The number of item identified as biased by each subsample is presented in Table 5. This table also presents the number of items which were identified in two, three or four of the subsamples.

Insert Table 5 about here.

DISCUSSION

The results suggest that the bias index based on the chi-square procedure is fairly unreliable when based on a sample size of 225 to 250 examinees in each ethnic group regardless of number of grouping intervals. All three measures of stability (correlations of chi-square values, correlations of 0,1 bias decisions, and the kappa index) indicated similar results. The correlations were varied but fluctuated from positive to negative values. The magnitude of the highest stability measure did suggest that the chi-square bias index could be expected to lead to consistent decisions about which items are biased.

Examination of the number of items identified as biased indicates that use of three ability groupings identifies a larger proportion of items as biased than do ability grouping of four or five. This larger number of items identified increased the number on which two or more subsamples agreed. However, in general, when two different samples were used to estimate bias, each sample identified relatively unique sets of biased items.

One potential explanation of the findings is that the English Usage test contained few, if any, biased items. During the test construction process, the items are screened for content validity as well as measurement properties. Items are also evaluated for potential bias by "judges." It could be that after such screenings, the chi-square statistical procedure, specially with smaller sample sizes, can detect little beyond what has already been evaluated by judges. This would suggest that the items identified were selected by chance factors. With an α level of .05, approximately four of 75 items would be expected to be identified by

chance. The number identified is slightly more than that for the four and five ability groupings and considerably more for the three ability groupings. Considering that the reliability of using three ability groupings is no higher than that for more ability grouping, the results may suggest that use of only three groupings introduces additional error into the bias detection index.

In the previous studies of reliability of item detection indices (Koler, et al., 1982; Loyd, 1982), five ability groupings were used in evaluating the chi-square procedure. The results of this study suggest that specifying fewer than five ability groupings would not have significantly altered the results which suggested that the index is fairly unreliable and cannot be expected to lead to consistent decisions about item bias.

REFERENCES

- Cohen, J. A. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.
- Ironson, G. H., & Subkoviak, M. J. A comparison of several methods of assessing item bias. Journal of Educational Measurement, 1979, 16, 209-225.
- Kolen, M. J., & Hoover, H. D. The Reliability of Selected Item Bias Procedures. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1982.
- Loyd, B. H. The Stability of Selected Item Bias Detection Methods. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1982.
- Rudner, L. M., Geston, P. R., & Knight, D. L. Biased item detection techniques. Journal of Educational Statistics, 1980, 5, 213-233.
- Scheunenman, J. A new method for assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.
- Shepard, L., Camilli, G., & Averill, M. Comparison of Six Procedures for Detecting Item Bias Using both Internal and External Ability Criteria. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, April 1980.

Table 1
Subsample Means, Standard Deviations
and Sample Sizes

	<u>Subsample</u>							
	1		2		3		4	
	B	W	B	W	B	W	B	W
Mean	32.7	46.4	34.3	46.3	33.4	45.7	35.1	45.4
Standard Deviation	11.9	12.6	12.2	12.4	11.2	12.6	12.3	12.8
N	222	243	247	265	221	241	231	249

Table 2

Correlations of Chi-square Indices for Subsamples
by Number of Ability Groupings

3 Ability Groupings

		Subsample			
		1	2	3	4
Subsample	1	-	.178	-.003	.381
	2		-	.188	.204
	3			-	-.126
	4				-

4 Ability Groupings

		Subsample			
		1	2	3	4
Subsample	1	-	.091	-.083	-.156
	2		-	-.126	.184
	3			-	.305
	4				-

5 Ability Groupings

		Subsamples			
		1	2	3	4
Subsample	1	-	.051	.053	.202
	2		-	-.068	-.048
	3			-	-.053
	4				-

All correlations calculated across 75 items.

Table 3

Correlations* of Bias Decisions for Subsamples
by Number of Ability Groupings

3 Ability Groupings

		Subsample			
		1	2	3	4
Subsample	1	-	.024	.085	.18
	2		-	.226	.069
	3			-	-.093
	4				-

4 Ability Groupings

		Subsample			
		1	2	3	4
Subsample	1	-	.186	-.076	.164
	2		-	-.082	.138
	3			-	-.088
	4				-

5 Ability Groupings

		Subsample			
		1	2	3	4
Subsample	1	-	.005	-.129	.220
	2		-	-.067	-.070
	3			-	-.070
	4				-

*Decisions coded 0,1 and correlations based on 75 items.

Table 4

Number of Items Identified as Biased by Two Subsamples
(Kappa Index of Agreement) for Subsamples by
Number of Ability Groupings

3 Ability Groupings

		Subsamples			
		1	2	3	4
Subsamples	1	-	3(.023)	2(.086)	4(.109)
	2		-	4(.207)	6(.163)
	3			-	0(-.172)
	4				-

4 Ability Groupings

		Subsamples			
		1	2	3	4
Subsamples	1	-	1(.005)	0(-.075)	2(.162)
	2		-	0(-.074)	2(.183)
	3			-	0(-.073)
	4				-

5 Ability Groupings

		Subsamples			
		1	2	3	4
Subsamples	1	-	1(.004)	0(-.121)	2(.185)
	2		-	0(-.188)	0(-.064)
	3			-	0(-.064)
	4				-

Table 5

Number of Item Identified as Biased

	By Subsample				By Two Subsamples	By Three Subsamples	By Four Subsamples
	1	2	3	4			
3 Ability Groupings	12	17	8	17	9	3	1
4 Ability Groupings	9	8	7	4	1	1	0
5 Ability Groupings	12	6	6	4	3	0	0